Università degli Studi di Padova

Facoltà di Scienze MM.FF.NN

Corso di Laurea in Biologia Molecolare

Elaborato di Laurea

Analisi *in silico* della proteina CDKL5

Tutor: Prof. Silvio Tosatto

Dipartimento di Biologia

Co-tutor: Dott.essa Emanuela Leonardi

Affiliazione: Dipartimento di Biologia

Laureando: Andrea Gazzo

Anno Accademico 2010/2011

Indice

Abstract

Questo elaborato di laurea descrive l'analisi *in silico* della proteina Cyclin-Dependent Kinase-Like 5 (CDKL5). CDKL5 è coinvolta in processi di sviluppo neuronale, infatti mutazioni che alterano la funzione di questa proteina portano come risultato fenotipico ad anomalie nel sistema nervoso. Dato che non esiste una struttura cristallizzata rappresentativa di questa proteina, per prima cosa è stato creato un modello tridimensionale putativo rappresentativo della stessa. Oltre alla suddivisione in subdomini e la ricerca di siti funzionali, ottenuti grazie ad un multiallienamento strutturale con altri domini chinasici meglio conosciuti, il modello è stato caratterizzato da vari punti di vista, come ad esempio la conservazione e la superficie elettrostatica. Inoltre è stato possibile effettuare mutagenesi *in silico* in maniera da capire come le varianti naturali per un singolo amminoacido riscontrate in natura siano differenti da un punto di vista strutturale rispetto alla proteina wild type. Per alcune di queste mutazioni è stato possibile dare una spiegazione di come queste differenze strutturali si riflettano in differenze a livello funzionale. Infine, lo studio della lunga coda C-terminale, della quale ancora poco è noto in letteratura, ne ha permesso un'iniziale caratterizzazione funzionale.

1 - Introduzione

1.1 CDKL5 malattie correlate

Il gene cyclin-dependent kinase-like 5 (cdkl5) è situato nel cromosoma X ed esprime una proteina coinvolta nello sviluppo neuronale. Delezioni o mutazioni amminoacidiche determinano la produzione di una proteina non funzionante o funzionante in maniera non corretta, le quali portano allo sviluppo di una malattia degenerativa spesso classificata in modo generico in sindromi come quella di Rett atipica, di West e di Lennox Gastaut [1,2]. I sintomi comuni, che si manifestano entro i primi sei mesi di vita, sono epilessia precoce generalmente non trattabile con farmaci, ritardo mentale ed autismo. Queste mutazioni si possono verificare sia in individui di sesso maschile sia femminile con una frequenza maggiore nelle donne. Non essendo riscontrata alterazione nei genitori è presumibile che la mutazione insorga "de novo" al momento del concepimento [2]. Le mutazioni che causano la malattia sono distribuite sia nel dominio catalitico che nel dominio C-terminale.

1.2 La proteina CDKL5

CDKL5 è coinvolta nello sviluppo neuronale. E' presente in quantità minime nello stato embrionale mentre è fortemente sovra-espressa nello sviluppo postnatale. In particolare la sua espressione viene indotta nei neuroni in maturazione della corteccia celebrale e dell'ippocampo. Da un punto di vista di localizzazione subcellulare CDKL5 è presente sia nel nucleo che nel citoplasma in concentrazioni diverse, che possono variare in base alla regione del cervello e allo stadio di sviluppo. Si suppone che la proteina faccia da spola tra un compartimento e l'altro. CDKL5 è costituita da 1.030 amminoacidi e fa parte della superfamiglia delle chinasi. Sono proteine che catalizzano il trasferimento di un gruppo fosfato da una molecola di ATP a un residuo di serina o treonina di proteine regolatrici e strutturali [3]. Per questo motivo CDKL5 veniva indicata con il nome alternativo di STK9, l'acronimo di Serine/threonine protein kinase 9. CDKL5 può essere suddivisa in due domini: il dominio catalitico, il quale consta dei primi 297 amminoacidi, e una lunga coda C-terminale, che comprende tutti gli amminoacidi rimanenti.

1.3 Le Proteine Chinasi

Le proteine chinasi possiedono dei motivi strutturali conservati, questi motivi forniscono indicazioni chiare su come questi enzimi riescono a trasferire il fosfato di un nucleotide trifosfato ai gruppi idrossilici delle loro proteine substrato.
Le chinasi che definiscono questo gruppo di enzimi contengono 12 sottodomini conservati, che si piegano in una struttura comune a tutte le chinasi [4], con la funzione di catalizzatore. Questi enzimi utilizzano il γ-fosfato di ATP (o GTP) e per generare monoesteri utilizzano come accettori di gruppi fosfato i gruppi alcolici (su Ser e Thr) e / o gruppi fenolici (su Tyr). Le chinasi agiscono secondo tre passaggi: l'orientamento del ATP come un complesso con catione bivalente (di solito Mg^{2+}); l'orientamento delle proteine (o peptidi) substrato; il trasferimento della γ -fosfato da ATP al residuo accettore ossidrile (Ser, Thr, Tyr) del substrato

proteico. La profonda spaccatura tra i due lobi è riconosciuto come il sito della catalisi. L'enzima viene inattivato mediante l'interazione di una sequenza peptidica autoinibitoria, che si trova sul lato COOH-terminale e si ripiega nella fessura attiva sito tra i due lobi (a volte può essere una molecola esterna, non solo un peptide dello stesso enzima). Questo peptide sembra forzare i due lobi a ruotare circa di 30° gradi l'uno rispetto all'altro (configurazione twitchin inattiva).
La nostra ricerca si concentra sul dominio chinasico che, nonostante conti meno di un terzo degli amminoacidi della proteina intera, rappresenta la parte funzionale nota della stessa. Il core catalitico è molto simile a quello di altre chinasi poiché, come ci si può immaginare, è altamente conservato [3]. Possiamo suddividere il dominio chinasico in due subdomini: quello N-terminale, più piccolo e ricco di foglietti beta e quello C-terminale, più grande e formato per la maggior parte da alfa eliche. Il core catalitico è situato all'interfaccia tra i due subdomini, i quali possono assumere una conformazione aperta o chiusa: questo dipende dal legame con l'ATP e dallo stato di attivazione della molecola.
Non molto è noto sul dominio C-Terminale. Si tratta di una lunga sequenza amminoacidica destrutturata, coinvolta nella localizzazione subcellulare e nell'autoregolazione dell'attività di CDKL5 [1,5]. Si ipotizza che alcuni amminoacidi della coda C-terminale interagiscano con il dominio catalitico mantenendolo in forma inattiva. Infatti i mutanti mancanti del dominio C-terminale o di una porzione di questo presentano un elevata attività chinasica, oltre che una localizzazione subcellulare anomala [5].

1.4 Interattori di CDKL5

Methyl CpG Binding Protein 2 (MeCP2) è una proteina esclusivamente nucleare che funge da repressore trascrizionale che agisce su diversi promotori. Si pensa che CDKL5 e MeCP2 siano coinvolte nello stesso pathway, a supporto di questa ipotesi ci sono evidenze sia cliniche che molecolari [1]. Per quanto riguarda le prime, mutazioni nel gene MeCP2 causano malattie simili a quelle causate da CDKL5 come ad esempio la sindrome di Rett. Da un punto di vista molecolare, CDKL5 e MeCP2 formano un complesso in vivo e l'attività catalitica di CDKL5 fosforila MeCP2 in vitro. Inoltre I due geni sono attivati simultaneamente durante lo sviluppo e condividono il pattern di espressione. CDKL5 ha anche altri target non conosciuti che portano ad uno specifico fenotipo associato [1].

1.5 Scopo della tesi

Numerosi dati sperimentali sono stati raccolti ma la funzione di questa proteina è ancora sconosciuta e, soprattutto, non esiste ancora una struttura cristallografica. Lo scopo della tesi è stato quello di fare una predizione della struttura tridimensionale del dominio chinasico di questa proteina e studiare su questo modello l'impatto strutturale e funzionale delle mutazioni identificate in pazienti con sindrome di Rett atipica o epilessia precoce riportate in letteratura. Per quanto riguarda il dominio C-terminale si è cercato di caratterizzare, anche se solo parzialmente, questa lunga sequenza con un elevato grado di disordine, cercando possibili siti funzionali all'interno di essa.

2 - Materiali e metodi

2.1 Analisi di Sequenza

La sequenza in formato fasta, punto di partenza per ogni ricerca computazionale di una proteina, è stata ottenuta dal database Swissprot. Oltre alla sequenza il database contiene informazioni di varia natura riguardo alle proteine: descrizione, localizzazione subcellulare, interattori, descrizione di mutazioni note, organizzazione in domini e sopratutto descrizione di motivi funzionali, come ad esempio zone di legame al DNA o ATP e siti di fosforilazione, permettendo una caratterizzazione iniziale della proteina a livello molecolare.
Spritz è un web server per la predizione della struttura secondaria e di regioni intrinsecamente disordinate in sequenze proteiche. Si può scegliere come modalità di computazione tra "short disorder", nel quale Spritz predirà più efficacemente regioni di disordine corte, e "long disorder", dove sarà più efficace la predizione di sequenze disordinate lunghe. L'output è composto da tre righe: la prima indica la sequenza analizzata , la seconda la predizione della struttura secondaria effettuata dal server Porter, la terza linea indica la predizione del disordine.
ELM è una risorsa per la predizione dei siti funzionali in proteine eucariotiche. I siti funzionali putativi sono identificati grazie ad espressioni regolari. Come input si inserisce la sequenza in formato fasta o l'ID di Swissprot della proteina. Parametri fondamentali per la corretta identificazione dei siti funzionali sono la specie e il compatimento cellulare dove la proteina è presente. L'output è costituito da una pagina web che indica, lungo tutta la sequenza, i siti funzionali putativi.
ANCHOR è un server che predice, da sequenze amminoacidiche con elevato grado di disordine, regioni che possono legare proteine. Queste regioni interagendo con una proteina possono subire cambiamenti conformazionali dal momento che gli amminoacidi che le compongono non sono vincolati in posizioni fisse. I metodi con i quali ANCHOR predice le possibili regioni di legame sono implementati con IUPred, un software per la previsione generale del disordine.

2.2 Analisi struttura terziaria

Le strutture tridimensionali di molte proteine vengono raccolte nel database Protein Data Bank (PDB). Queste strutture sono state ottenute tramite cristallografia a raggi X o NMR. Il file che rappresenta ogni proteina contiene le coordinate spaziali degli atomi di tutti gli amminoacidi che compongono la struttura. E' importante sottolineare il fatto che la gran parte delle strutture ottenute tramite cristallografia a raggi X provengono da proteine che al momento della diffrazione erano legate ad un determinato ligando, a ciò consegue il fatto che la struttura del file PDB rappresenta lo stato conformazionale della proteina legata a quel ligando, che non sempre rappresenta la conformazione della proteina funzionalmente attiva.
CE è un software che crea allineamenti strutturali: confronta varie catene polipeptidiche usando come criterio la loro geometria locale, cioè effettua un multiallineamento in base agli elementi di struttura secondaria presenti nei vari peptidi che partecipano al multiallineamento.

PyMOL è un software di grafica 3D utilizzato per la rappresentazione di biomolecole in bioinformatica. Si tratta di un software open-source ed è in grado di girare su sistemi operativi diversi, la vasta gamma di opzioni permette di visualizzare strutture molecolari, come proteine, focalizzandosi su aspetti diversi.

2.3 Creazione Modello e Mutagenesi in silico

HOMER (HOmology ModellER) è un server di modellazione comparativa per la predizione di strutture proteiche. In modalità di selezione automatica si inserisce come input una sequenza in formato fasta, HOMER produrrà come output una lista di possibili templati indicando per ognuno Bit Score, grado di conservazione ed e-value. In modalità di selezione manuale del templato si fornisce l'allineamento tra la sequenza della proteina di cui si vuole costruire il modello e la sequenza della proteina che si vuole usare come templato e oltre a questo la struttura PDB della proteina che si vuole utilizzare come templato. Il programma fornisce in output un file contenente le coordinate atomiche del modello e una valutazione energetica della struttura così ottenuta. Homer è stato utilizzato inoltre per la costruzione dei modelli delle strutture tridimensionali dei mutanti utilizzando come templato il modello wild type della proteina precedentemente creato. Tutti I modelli ottenuti sono stati sottoposti a minimizzazione energetica utilizzando il programma Gromacs. In seguito grazie al programma QMEAN si è potuta fare una stima della qualità dei modelli ottenuti.

2.4 Analisi di conservazione e superficie elettrostatica

Il server ConSurf è uno utile strumento che consente l'identificazione di regioni funzionalmente importanti sulla superficie di una proteina o di un dominio, partendo dalla struttura tridimensionale, sulla base delle relazioni filogenetiche tra le sequenze degli omologhi più vicini alla proteina data. Le sequenze omologhe sono state selezionate automaticamente da Consurf. La relazione tra conservazione e rilevanza funzionale è data dalla considerazione che i residui funzionalmente importanti con molta probabilità sono conservati nel corso dell'evoluzione.
Utilizzando il plug-in di PyMol APBS si può analizzare la superficie elettrostatica della proteina. Per prima cosa di ottiene un file PQR inserendo come input il file PDB della proteina di interesse nel server http://kryptonite.nbcr.net/pdb2pqr/. Il file PQR è un file PDB comprendente il profilo elettrostatico di ogni amminoacido. Il file PQR viene visualizzato grazie a APBS che permette di focalizzarsi sulle differenze di carica elettrostatica superficiale grazie a colorazioni diverse.

2.5 Metodi per l'analisi delle mutazioni missenso e dei loro effetti

Gli studi sperimentali sugli effetti molecolari delle mutazioni sono molto laboriosi, tuttavia i metodi di predizione bioinformatica possono indirizzare la strategia sperimentale da impiegare. Pon-P è un portale che mira a fornire una vasta collezione di strumenti bioinformatici per l'analisi dell'effetto delle mutazioni. Da questo portale sono stati scelti alcuni fra i metodi per la predizione della stabilità della proteina mutante.
I-Mutant 2.0 è un programma in grado di valutare il grado di cambiamento di

stabilità sulla struttura causato dalla mutazione di un singolo amminoacido a partire dalla struttura o dalla sequenza della proteina. Si basa su valori ottenuti sperimentalmente e raccolti nel database ProTerm. I-Mutant2.0 predice correttamente se una mutazione stabilizza o destabilizza la struttura di una proteina con l'80% dell'efficacia quando la struttura tridimensionale è nota e con il 77% nel caso in cui solo la sequenza è disponibile. Si possono impostare come parametri la temperatura e il Ph. Nel mio caso ho utilizzato quelli impostati di default, rispettivamente 25 e 7. L'output fornito indica un valore di $\Delta\Delta G$ tra il ripiegamento della proteina wild type e quello della proteina mutante. Se $\Delta\Delta G$ è negativo la mutazione è destabilizzante, al contrario se è positivo è stabilizzante.
SCide è un programma atto a identificare i centri di stabilizzazione nelle proteine: si tratta di gruppi di residui coinvolti nelle interazioni a lungo raggio. A causa della loro natura cooperativa, queste interazioni svolgono un ruolo chiave nel mantenere la stabilità delle strutture proteiche. L'input è costituito dall'ID del PDB o da un file in formato .pdb direttamente caricato dall'utente.
SCPRED è un programma che utilizza una rete neurale per la predizione dei residui coinvolti in interazioni forti a lungo raggio. Come SCide questo programma serve a identificare i centri di stabilizzazione ma per fare ciò si serve solo della sequenza amminoacidica. L'output contiene due righe: la prima è la sequenza, la seconda contiene le cifre '0' o '1'. '0' indica i residui non coinvolti in centri di stabilizzazione, '1' quelli coinvolti.
MUPRO è un insieme di programmi per predire come un unico sito mutazione amminoacidica possa influenzare la stabilità di una proteina. E' basato su due metodi di machine learning: Support Vector Machines e reti neurali. Si può inserire come input solo la sequenza o sia sequenza che la struttura. Naturalmente, se si fornisce la struttura terziaria, si ottiene una predizione migliore. Oltre alla sequenza ed eventualmente la struttura l'input è costituito anche dalla posizione e dalla natura della mutazione. L'output consiste in una predizione di stabilità che può risultare incrementata o decrementata dalla mutazione.

2.6 Analisi reti di interazioni

RING (Residue Interaction Network Generator) è un web server per trasformare una struttura proteica in una rete di interazioni. I nodi rappresentano i singoli amminoacidi nella struttura della proteina, mentre gli archi rappresentano le interazioni non covalenti che esistono tra loro. RING è in grado di elaborare un singolo file PDB e produce come output la rete di interazione e gli attributi dei nodi e delle interazioni. Questi file possono essere facilmente caricati in CYTOSCAPE, programma grazie al quale è possibile visualizzare e manipolare la rete. Gli attributi dei nodi e degli archi sono numerosi, quelli che abbiamo analizzato più da vicino sono: per quanto riguarda i nodi il grado di accessibilità al solvente e il grado di conservazione, per quanto riguarda gli archi la natura del legame e le distanze tra gli amminoacidi che interagiscono.
Usando come input per RING diversi PDB che rappresentano il wild type e i mutanti puntiformi, è possibile creare diverse reti rappresentative.

3 - Risultati

3.1 Analisi sequenza

La porzione N-terminale di CDKL5 è riconosciuta dal Database dei Domini Conservati (CDD) come dominio catalitico appartenente alla famiglia delle proteine chinasi Ciclin-dipendent (CDKs). A loro volta, le CDKs fanno parte della superfamiglia delle Protein Chinasi, e hanno la peculiarità di poter interagire con le cicline, le quali regolano l'attività delle chinasi a cui si legano. I complessi ciclina-CDK sono coinvolti nel controllo del ciclo cellulare, trascrizione e sviluppo neuronale. Utilizzando BLAST per la ricerca di sequenze omologhe a CDKL5 si ottiene una lista di proteine chinasi, che comprende oltre alle CDKs anche le Mitogen Activated Protein (MAPKs), le chinasi regolate da segnali extracellulari (ERKs), le c-Jun N-terminal chinasi (JNKs), le glicogeno sintetasi chinasi (GSKs), le CDK-like chinasi(CLKs), e proteine simili. Un dato interessante ottenuto dall'analisi della sequenza riguarda il motivo "TEY", un sito di fosforilazione appartenente al dominio chinasico il quale ha un ruolo importante nel processo di attivazione della proteina. La predizione della struttura secondaria del loop nel quale è presente il motivo "TEY" indica che è disordinato, questo è in accordo con la posizione e il ruolo funzionale del loop: uno stato disordinato rende la struttura maggiormente plasmabile alle modificazioni indotte da stress ambientali o alle interazioni coi ligandi. La fosforilazione della tirosina provoca un cambiamento conformazionale del loop con conseguente modificazione della struttura che a sua volta si riflette in un cambiamento funzionale cioè il passaggio dallo stato inattivo a quello attivo.

3.2 Analisi struttura 3D

Il modello della struttura di CDKL5 è stato creato utilizzando come templato 2BKZ, una struttura cristallizzata depositata su Protein Data Bank rappresentativa della proteina CDK2. Homer fornisce una lista di possibili templati in cui la struttura del dominio chinasico della Cyclin-dependent kinase 5 (PDB ID: 1UNGA) risulta avere il bit score più elevato. Tuttavia l'analisi delle diverse strutture da utilizzare come templato ha messo in evidenza la presenza, in molte di queste, di piccole regioni disordinate non presenti nel cristallo. Questo avrebbe influito nella costruzione del modello per cui è stata scelta la struttura di CDK2 (PDB ID: 2BKZ) che è completamente cristallizzata. La struttura cristallizzata di 2BKZ, presa come modello, è in complesso con la ciclina A2 e l'inibitore SBC.
La valutazione della qualità del modello è stata effettuata con QMEAN (***Figura 1***) e ci ha permesso di distinguere tre regioni di bassa qualità: il loop che segue il foglietto β3 nel subdominio SDII, che corrisponde al loop catalitico; il loop dopo il foglietto β9 nel subdominio SDVIII che contiene il motivo "TEY" e il loop che connette le α-eliche G e H nel subdominio SDX. Nel resto della struttura del modello la qualità è buona, sopratutto nelle zone attive nel ruolo catalitico e per quelle di legame all'ATP. Solitamente i domini chinasici hanno un ripiegamento simile, caratterizzato da motivi funzionali comuni. Per l'identificazione dei subdomini è stato preso come riferimento il modello sulla PKA come struttura rappresentativa delle proteine chinasi in organismi eucariotici [4].

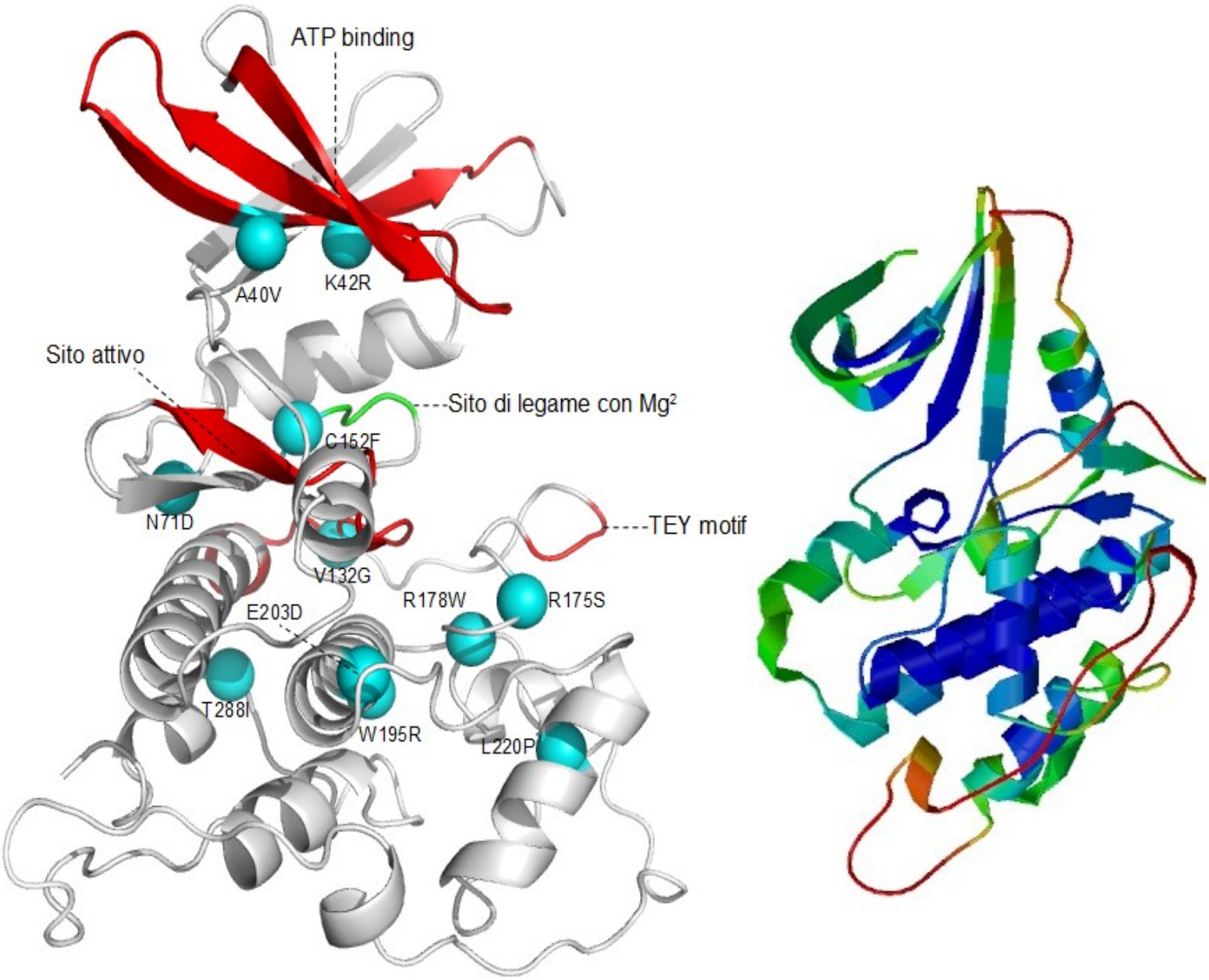

***Figura 1**: A sinista, rappresentazione del modello tridimensionale con mappate alcune tra le più importanti mutazioni. Sono indicati anche i motivi funzionali. Nell'immagine a destra è rappresentata un analisi della qualità del modello effettuata con Qmean. Blu: alta qualità, Rosso:bassa qualità*

L'immagine a sinistra nella ***Figura 1*** è la rappresentazione del modello della struttura trimensionale ottenuta con Homer. E' possibile individuare I motivi funzionali noti della proteina, oltre a questo una panoramica delle mutazioni più importanti sotto forma di sfere.

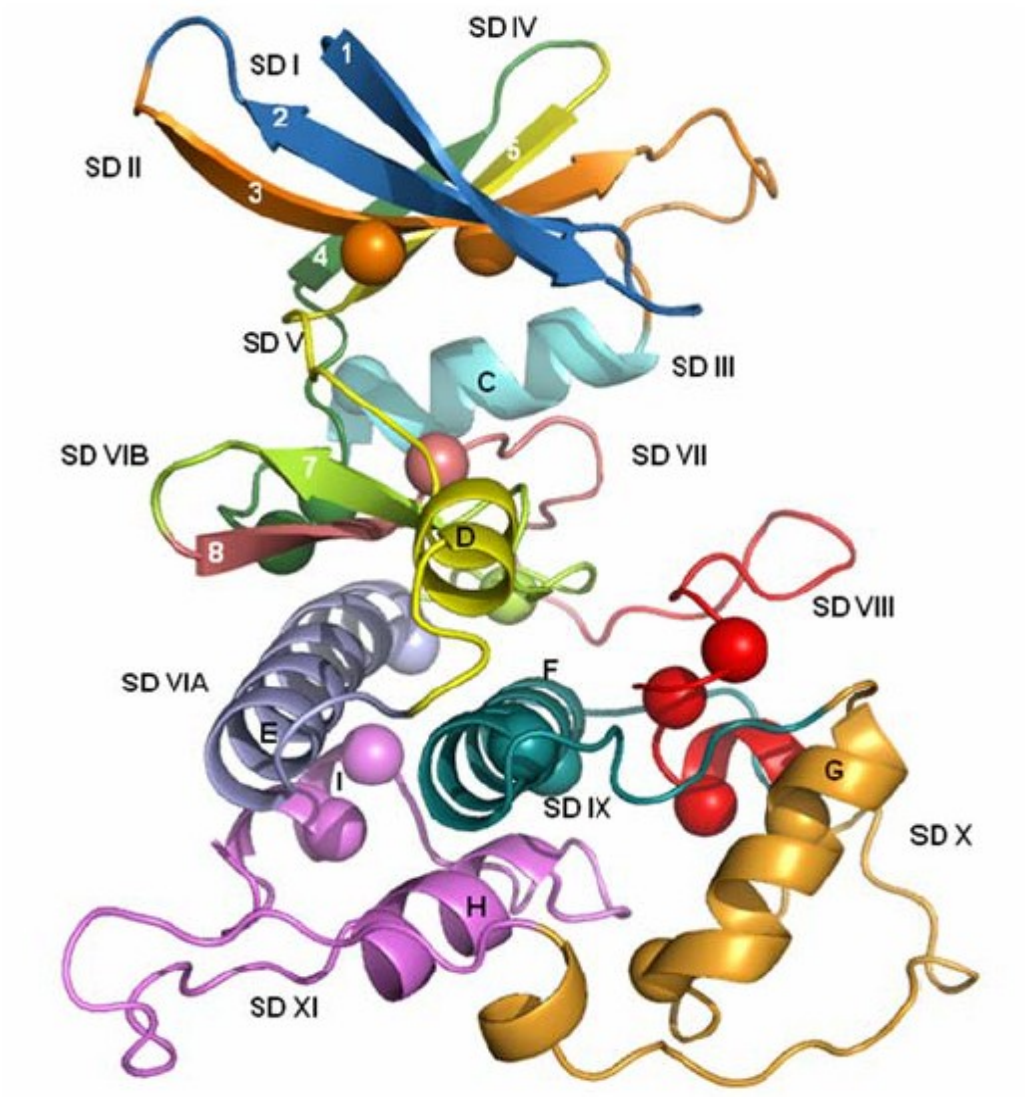

Figura 2 *: Un allineamento strutturale tra PKA e CDKL5 ha permesso di definire i 12 subdomini. La figura mostra il modello di CDKL5 colorato in base ai subdomini. Le lettere e i numeri indicano gli elementi di struttura secondaria. Compaiono mappate alcune delle mutazioni più importanti, presenti sotto forma di sfere.*

Analizzando i dati della ***Tabella 1*** è possibile notare come molte delle mutazioni siano localizzate in motivi funzionali del dominio chinasico (***Figura 2***).
Questo dato è in accordo con il fatto che mutazioni puntiformi situate all'interno di regioni funzionali producano un malfunzionamento della proteina con conseguente sviluppo di un sistema neuronale anomalo, riscontrato nei pazienti che presentano queste mutazioni.

SD	Range	Struttura II	Motivo funz.	Mutazioni
I	13 - 34	b1 - b2	G - loop	
II	35 - 54	b3	catalitic loop	A40V, K42R
III	55 - 67	C		R65Q
IV	68 - 82	b4		N71D, I72R, I72N
V	83 - 105	b5 - D	linker region	
VIA	106 - 129	E		H127R
VIB	130 - 146	b6 - b7	active site	V132G
VII	147 - 162	b8 - b9	Metal binding	C152F
VIII	163(167) - 183	α - helix	activation loop	R175S, R178P, R178W, P180L
IX	184 - 213	F		W195R, E203D
X	214(233) - 252	G - a - a - a		L220P, L227R
XI	253(266) - 297	H - I		T288I, C291Y

Tabella 1: *viene fornito un profilo topografico di tutte le mutazioni analizzate*

Dall'analisi della superficie elettrostatica si evince che il dominio catalitico presenta due regioni di carica opposta (***Figura 3***). La regione carica positivamente, mappa a livello del lobo superiore, che corrisponde alla zona di legame con l'ATP. Questo è in accordo con il fatto che I gruppi fosfati con I quali deve interagire sono carichi negativamente. La regione carica negativamente si trova invece nel lobo inferiore e nella zona del lobo superiore limitrofa all'interfaccia tra I due subdomini. Questa regione coincide con la zona di legame al substrato e di conseguenza ci si può aspettare che la parte di substrato che interagisce in maniera diretta con la proteina potrebbe essere carica positivamente. L'analisi della conservazione della rivela la presenza di un core proteico conservato e di una superficie conservata all'interfaccia tra i due lobi (***Figura 3***). E' normale che il nucleo catalitico di un enzima risulti fortemente conservato: nel corso dell'evoluzione gli amminoacidi della zona funzionalmente attiva della proteina tendono a rimanere gli stessi e hanno una probabilità di essere sostituiti nettamente minore rispetto a quelli delle zone periferiche, non direttamente coinvolti nella funzione catalitica dell'enzima. Per quanto riguarda la zona di legame con l'ATP è localmente meno conservata, però alcuni amminoacidi in particolare risultano essere molto conservati. Gli amminoacidi che contattano direttamente l'ATP risultano i più conservati, mentre quelli nelle vicinanze, anche se nel loro complesso mantengono una carica nettamente positiva, siano stati sostituiti nel corso dell'evoluzione.

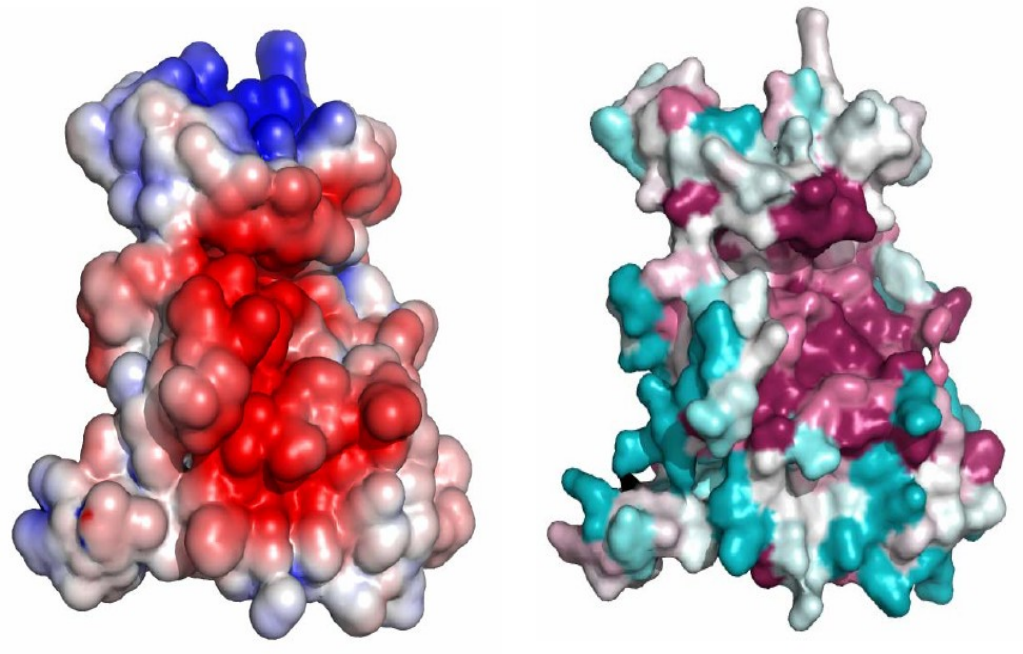

Figura 3: *La figura a sinistra mostra la superficie elettrostatica del dominio chinasico di CDKL5. La colorazione blu indica carica positiva, quella rossa carica negativa. La figura a destra indica la superficie di conservazione del dominio chinasico di CDKL5. La colorazione Magenta indica elevato grado di conservazione, Cyan indica basso grado di conservazione.*

3.3 Analisi degli effetti delle mutazioni

Mutazioni di singoli amminoacidi possono modificare la struttura tridimensionale di una proteina alterando anche la funzionalità della stessa. Esistono molti metodi computazionali per predire l'effetto di una mutazione sulla struttura tridimensionale. Per l'analisi delle mutazioni del dominio chinasico di CDKL5 sono stati utilizzati diversi programmi di predizione della stabilità. Molti di questi programmi calcolano lo stesso tipo di predizione ma usano algoritmi diversi per raggiungere il risultato finale. Ottenere un consensus fra metodi diversi permette di consolidare i risultati. La ***Tabella 1A*** in appendice mostra diverse predizioni: SCide e SCPRED predicono i centri di stabilità, la differenza è che nel primo caso viene usato come input la struttura nel secondo la sequenza. I-Mutant 2.0 quantifica con un valore di DDG il grado di destabilizzazione dovuto alla mutazione. Il grado di conservazione e di accessibilità al solvente sono stati ricavati dagli attributi dei nodi ottenuti grazie a RING. RING inoltre permette una analisi qualitativa sui cambiamenti a livello locale nella rete di interazioni che rappresenta il dominio chinasico della proteina. Mutazioni in singoli amminoacidi possono avere effetti diversi sui legami che il nuovo amminoacido stringe con gli amminoacidi limitrofi, questo dipende principalmente dalle caratteristiche del nuovo amminoacido e dalla regione in cui avviene la mutazione. Abbiamo scelto tre parametri per classificare il grado di cambiamento dei legami nei mutanti con RING: "poco significativo" nel caso in cui il cambiamento di legami sia limitato a leggere variazioni di distanza o la perdita o modifica di uno o due legami; "significativo" nel caso di cambiamenti della rete che compromettono numerosi

legami, “drastico” nel caso in cui la maggior parte della rete di interazioni risulti compromessa dalla mutazione amminoacidica.

3.4 Analisi del dominio C-terminale

La maggior parte di amminoacidi che compongono CDKL5 non fanno parte del dominio chinasico e formando una lunga coda C-terminale. La predizione di Spritz per quanto riguarda il dominio C-terminale risulta altamente disordinata, tuttavia sono presenti alcuni elementi di struttura secondaria. La predizione di ELM per quanto riguarda la ricerca di siti funzionali, mostra numerosi siti di riconoscimento per l'import e l'export nucleare, dato in accordo con l'ipotesi che la proteina CDKL5 possa fungere da shuttle tra citoplasma e nucleo [5]. Nonostante il filtro impostato per la limitazione della ricerca ai siti funzionali delle sole proteine nucleari(CDKL5 è presente maggiormente nel nucleo), lungo tutta la sequenza della coda C-terminale sono presenti numerosissimi siti funzionali putativi riconosciuti da ELM. Per riuscire a filtrare I dati più interessanti provenienti da ELM abbiamo utilizzato un altro software, ANCHOR, il quale è in grado di predire siti di legame in sequenze amminoacidiche disordinate. La ***Figura 4*** mostra l'immagine della predizione.

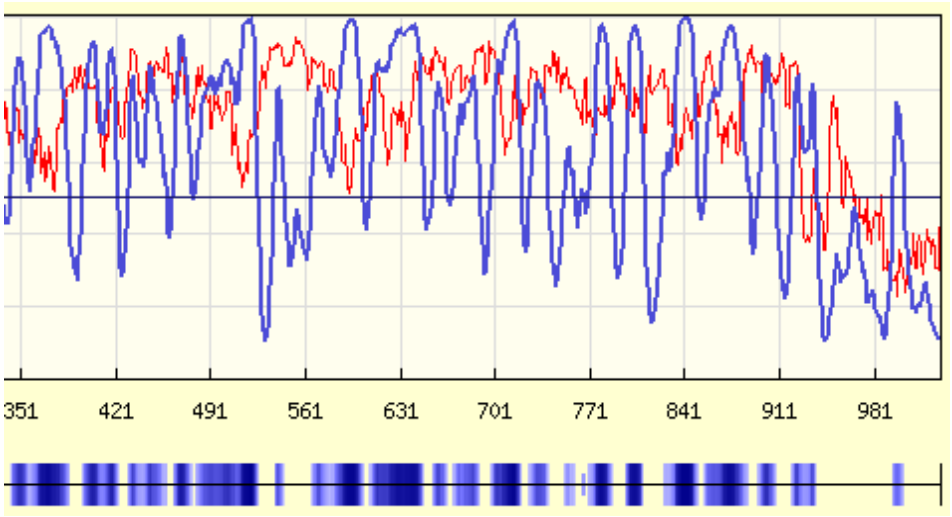

Figura 4: *rappresentazione delle zone di legame della coda C-Terminale calcolate da ANCHOR. Ancor predice oltre ad un elevato grado di disordine (picchi rossi, calcolati con IUPred), molte regioni disordinate di possibile interazione con altre proteine (picchi blu).*

Per analizzare I dati significativi abbiamo tenuto conto di tre fattori: regioni disordinate con propensione a formare legami con proteine (ANCHOR), siti di fosforilazione noti (UniProt) e proteine che interagiscono in subregioni, all'interno delle regioni di legame, che coprono anche il sito di fosforilazione noto. I dati sono stati riassunti in ***Tabella 2***.

Regione di Legame	Sito di fosforilazione	Interattori predetti
344 –386	375 e 377	GSK3, MAPK,Class IV WW
397 - 423	407	Class IV WW*, CK2, MAPK
464 - 478	476	GSK3
480-526	488	GSK3
609-649	646	GSK3,Class IV WW
670-690	681	CK1,GSK3
699-721	720	Class IV WW, MAPK

Tabella 2: *ogni colonna rappresenta, rispettivamente la regione di legame predetta da ANCHOR, il sito di fosforilazione noto presente su Uniprot, e gli interattori predetti da Elm all'interno dei siti di legame di ANCHOR.*

I Class IV WW domains interaction motifs sono un insieme di motivi di interazione dipendenti da fosforilazione, sono motivi implicati nelle interazioni mediate da domini WW. Questi domini sono lunghi da 38 a 40 residui e mediano l'interazione tra proteine attraverso il legame con brevi regioni ricche in prolina. Le proteine che presentano questi domini sono coinvolte in molti processi cellulari, tra cui la degradazione mediante ubiquitinazione e regolazione mitotica.

4 - Discussione

I dati ottenuti grazie a strumenti di bioinformatica spesso hanno bisogno di essere affiancati ai dati provenienti da esperimenti in vitro o in vivo. L'interazione tra queste diverse branche di ricerca della biologia molecolare è necessaria per poter comprendere risultati che altrimenti risulterebbero essere solo provenienti da metodi computazionali e quindi non verificabili. In questo capitolo discuteremo in maniera critica I risultati, mettendoli, quando possibile, in relazione con I dati provenienti dalla ricerca in "wet lab".

4.1 Riflessione sui cambiamenti strutturali delle mutazioni

Da un analisi della ***Tabella 1A*** in appendice, emerge che il quadro descrittivo delle mutazioni è destabilizzante da un punto di vista strutturale per tutte. Questo dato non ci sorprende dal momento che si tratta di mutazioni riscontrate in pazienti che hanno sviluppato un sistema nervoso compromesso. Le mutazioni A40V e K42R di per se non hanno un impatto strutturale devastante, oltre ai dati dell'analisi di RING, SCide e SCPRED indicano che sono centri di stabilità che vengono mantenuti. Tuttavia, come indica il grado di conservazione elevatissimo, si trovano in un punto chiave: la zona di legame all'ATP, per cui entrambe le mutazioni portano alla perdita dell'attività chinasica. Tutte le mutazioni analizzate sono interne, tranne R65Q e R175S che presentano un elevato grado di accessibilità al solvente. La prima di queste due mutazioni sembra essere quella meno dannosa in tabella, tuttavia non avendo dati biochimici sull'attività di questa proteina mutante non è possibile fare inferenze. Nel secondo caso invece la mutazione sembra creare differenze strutturali, rispetto al wild type, molto pronunciate, e dal momento che è localizzata nel loop di attivazione sembra chiaro il motivo per il quale la proteina mutata sia priva di attività chinasica; inoltre tra tutte le mutazioni è quella che risulta più destabilizzata secondo i valori di I-Mutant2.0. R175S, R178P, R178W e P180L mappano tutte nel loop di attivazione, in prossimità del motivo "TEY", non sorprende quindi il fatto che tutte abbiano grado di conservazione massimo e la loro sostituzione possa portare alla perdita dell'attività chinasica. Il subdominio IX è quello che dovrebbe interagire con il peptide inibitore (un tratto della coda C-terminale), in questo subdominio sono mappate le mutazioni W195R e E203D: le due mutazioni hanno un grado di destabilizzazione predetto da I-Mutant 2.0 nettamente diverso, questo probabilmente è in funzione al tipo di mutazione amminoacidica. In W195R il triptofano (W) è idrofilico e solo leggermente polare, mentre l'arginina (R) con il quale viene sostituito è un amminoacido fortemente basico. Nel caso di E203D l'acido glutammico (E) e l'acido aspartico (D) presentano una medesima carica, ciò potrebbe in parte spiegare il valore molto basso calcolato da I-Mutant 2.0. La mutazione L220P porta alla perdita di attività chinasica [6]: in effetti il grado di conservazione è massimo e SCide predice questa mutazione come causa della perdita del centro di stabilizzazione. Tuttavia questo dato non è in accordo con SCPRED e con la nostra analisi con RING; è importante evidenziare che il grado di destabilizzazione predetto da I-Mutant 2.0 è decisamente alto e questo fa propendere verso l'ipotesi che la mutazione sia altamente destabilizzante. Inoltre è presente un altro fattore da tenere in considerazione. La mutazione mappa all'interno di un alfa-elica (G): è possibile che la prolina, amminoacido che

sostituisce la leucina, destabilizzi fortemente questo elemento di struttura secondaria.

4.2 Ipotesi per la localizzazione subcellulare di CDKL5

In letteratura è stato ipotizzato che la localizzazione subcellulare fosse correlata all'attività chinasica, però analizzando i dati inerenti le concentrazioni di forme mutanti di CDKL5 [6] denota che sia il mutante C152F che K42R sono apparentemente privi di attività chinasica. Entrambi hanno una localizzazione subcellulare anomala ma tuttavia diversa tra loro. C152F ha una concentrazione superiore nel citoplasma rispetto al nucleo, a differenza del wild type [1]. Questo vale anche K42N, ma la sua concentrazione citoplasmatica è significativamente maggiore rispetto a C152F. Se la localizzazione subcellulare fosse direttamente correlata all'attività chinasica ci si aspetterebbe una localizzazione uguale nei due casi, invece non è propriamente cosi. Gli autori indicano come possibile motivazione il fatto che forse C152F potrebbe avere una leggera attività chinasica, troppo bassa per essere rilevata, mentre K42N no. Tuttavia dallo nostro studio sulla stabilità risulta che la mutazione C152F è molto destabilizzante da un punto di vista strutturale e dal momento che si trova in un punto cardine, cioè all'interfaccia tra lobo superiore e lobo inferiore dove è presente la tasca catalitica, per di più a ridosso del sito di legame al magnesio, sembra molto probabile che effettivamente la proteina mutata non possa svolgere la sua funzione chinasica. Un ipotesi alternativa potrebbe essere che la localizzazione sia correlata alla capacità di legare ATP. La mutazione K42R è localizzata nel sito di legame con l'ATP, strutturalmente la mutazione non è molto destabilizzante (***Tabella 1A*** in appendice). Dato che il core catalitico rimane praticamente intatto nei mutanti K42R, la mancanza di attività chinasica potrebbe essere attribuibile all'impossibilità di legare ATP. Al contrario C152F ha il nucleo catalitico molto destabilizzato ma il sito di legame all'ATP resta invariato rispetto al wild type, quindi è possibile che questo mutante possa legare ATP anche se in concentrazione minore, e questo determinerebbe una concentrazione subcellulare intermedia tra il wild type e il mutante K42R, cosa che effettivamente risulta dai dati sperimentali [1].

4.3 Possibili interattori

Il ruolo biologico della lunga coda C-terminale ha dato luogo a numerose ricerche, l'unica cosa che si compreso per certo è che ha attività autoinibitoria sul dominio catalitico [5]. Ma oltre a questo sicuramente molte altre funzioni possono essere attribuite al dominio C-terminale, come la possibilità di fungere da shuttle per permettere l'import-export nucleare di molecole e l'influenza sulle concentrazioni subcellulari di CDKL5. Dai dati che abbiamo ottenuto in ***Tabella 2*** possiamo inferire che I promotori della fosforilazione dei siti noti (Uniprot) sono protein chinasi appartenenti alle famiglie GSK3, CK2, MAPK, CK1. Questo elenco presenta famiglie di protein chinasi legate alla regolazione del ciclo cellulare, dato coerente con la funzione nota di CDKL5. Su UniProt sono presenti informazioni riguardo a variazioni amminoacidiche naturali: la proteina con una mutazione in posizione 374 è stata riscontrata in campioni ottenuti da melanoma in fase di metastasi, sotto forma di mutazione somatica. Dal momento che questa

variazione è all'interno di regioni funzionali presenti in tabella, l'ipotesi della presenza di un potenziale motivo funzionale in questa regione sembra plausibile. Lo stesso vale per la variazione naturale in posizione 718, che è stata riscontrata in pazienti affetti da encefalopatia epilettica infantile precoce di tipo 2. Nelle regioni che presentano più di una protein chinasi come possibile interattore, probabilmente solo uno di quelli elencati è quello corretto, ma per quanto riguarda le regioni che presentano come possibili interattori sia protein chinasi che interattori della classe IV WW ci si aspetta sia il riconoscimento da parte di proteine chinasi che il riconoscimento da parte di proteine di questa classe, poiché gli interattori della classe IV WW sono dipendenti dalla fosforilazione delle regioni che riconoscono. Infine, un elemento interessante è stato riscontrato all'interno delle regioni disordinate di legame predette da ANCHOR, prima che si filtrassero i risultati selezionando solo I motivi funzionali in prossimità di siti di fosforilazione noti. Il motivo funzionale di legame a P53 compare numerose volte, per la precisione ci sono 10 siti funzionali di legame a P53 putativi all'interno delle regioni predette da ANCHOR.

Bibliografia

1) Ilaria Bertani, Laura Rusconi, Fabrizio Bolognese, Greta Forlani, Barbara Conca, Lucia De Monte, Gianfranco Badaracco, Nicoletta Landsberger, and Charlotte Kilstrup-Nielsen; *Functional Consequences of Mutations in CDKL5, an X-linked Gene Involved in Infantile Spasms and Mental Retardation;* **Journal of Biological Chemistry**; 2006.

2) **www.cdkl5.org**

3) Eric D.Scheeff, Philip E. Bourne; *Structural Evolution of the Protein Kinase–Like Superfamily;* **PLoS Computational Biology**; 2005.

4) Steven K. Hanks and Tony Hunter; *The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification*; **The FASEB journal**; 1995

5) Laura Rusconi, Lisa Salvatoni, Laura Giudici, Ilaria Bertani, Charlotte Kilstrup-Nielsen, Vania Broccoli, Nicoletta Landsberger; *CDKL5 expression is modulated during neuronal development and its sub-cellular distribution is tightly regulated by the C-terminal tail;* **Journal of Biological Chemistry**; 2008.

6) Nadia Bahi-Buisson, Juliette Nectoux, Haydeé Rosas-Vargas, Mathieu Milh, Nathalie Boddaert, Benoit Girard, Claude Cances, Dorothée Ville, Alexandra Afenjar, Marlène Rio, Delphine Héron, Marie Ange N'GuyenMorel, Alexis Arzimanoglou, Christophe Philippe, Philippe Jonveaux, Jamel Chelly and Thierry Bienvenu; *Key clinical features to identify girls with CDKL5 mutations;* **Brain**; 2008.

Mutazioni	Subdominio	I-Mut.	MUPRO	SCide	SCPRED	Con	SA	RING
A40V	SDII, ATP binding	-1,01	destabilizzante	SC M	SC M	9	2 (2)	Poco significativa
K42R	SDII, ATP binding	-0,82	destabilizzante	SC M	SC M	9	6 (6)	Poco significativa
R65Q	SDIII	-0,12	destabilizzante	-	-	7	54 (60)	Poco significativa
N71D	SDIV	-0,65	destabilizzante	SC M	-	9	2 (4)	Poco significativa
I72R	SDIV	-1,56	destabilizzante	SC M	SC P	7	0 (1)	
I72N	SDIV	-0,49	destabilizzante	SC M	SC P	7	1 (1)	
H127R	SDVIA	-0,03	destabilizzante	-	SC M	9	4 (2)	Poco Significativa
V132G	SDIVB, sito attivo	-1,21	destabilizzante	SC M	SC M	6	0 (0)	Significativa
C152F	SDVII, sito attivo	-0,97	destabilizzante	-	-	9	1 (0)	Drastica
R175S	SDVIII	-3,54	destabilizzante	-	-	9	25 (40)	Drastica
R178W	SDVIII	-1,11	destabilizzante	-	-	9	4 (6)	Poco Significativa
R178P	SDVIII	-0,41	destabilizzante	SC P	-	9	2 (6)	Significativa
P180L	SDVIII	-1,89	destabilizzante	SC M	-	9	0 (0)	Poco Significativa
W195R	SDIX	-2,26	destabilizzante	-	SC M	9	2 (0)	Significativa
E203D	SDIX	-0,01	destabilizzante	-	SC M	9	1 (2)	Significativa
L220P	SDX	-1,46	destabilizzante	SC P	SC M	9	1 (3)	Poco significativa
L227R	SDX	-1,27	destabilizzante	-	SC P	5	2 (3)	Significativa
T288I	SDXI	-0,77	destabilizzante	-	SC N	7	0 (1)	Significativa
C291Y	SDXI	0,35	destabilizzante	-	-	8	0 (0)	Significativa

Tabella 1A: *riassunto dell'impatto strutturale delle mutazioni. SC: centro di stabilità, M: mantenuto anche nel mutante, P: perso nel mutante, N nuovo nel mutante. I-Mut: I-Mutant2.0, SA: accessibilità al solvente calcolata con RING del mutante, (accessibilità al solvente nel wild type), Con: grado di conservazione calcolato con RING.*